Assignment 8 Submission

CS 432

Timothy Filippone

1. Create two datasets; the first called Testing, the second called Training.

Using my email tfili001@odu.edu, I copied two sets of ten inbox emails and another two sets from my spam folder. The spam folder did contain Twitter ads, which I do consider spam. The data sets were put into four text files, then converted to .db files using `termsql`.

```
termsql -c nameOfText,contents -i TrainingSpam.txt -o TrainingSpam.db
```

2. Using the PCI book modified docclass.py code and test.py (see Slack assignment-8 channel)
Use your Training dataset to train the Naive Bayes classifier (e.g., docclass.spamTrain())
Use your Testing dataset to test (test.py) the Naive Bayes classifier and report the classification results.

Both the default classifiers and ones containing “twitter” and “follow” were set in `docclass.py`. Neither configurations detected spam in the text content, even with twitter content being visible in the spam datasets. These results may have come from my personal strict set of spam filtering on my email account.

```
Classifying "the banking dinner" as Spam
           Spam  Not Spam
Training              X
Training Spam         X
Testing               X
Testing  Spam         X
----------------------------------------
Classifying "twitter" and "follow" as Spam
           Spam  Not Spam
Training              X
Training Spam         X
Testing               X
Testing  Spam         X
```